DOCUMENT RESUME

ED 249 287 TM 840 646

AUTHOR	Powers, Donald E.
TITLE	Considerations for Developing Measures of Speaking and Listening. College Board Report.
INSTITUTION	College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J.
REPORT NO	CB-R-84-5; ETS-RR-84-18
PUB DATE	84
NOTE	14p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101 ($5.00).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Cost Effectiveness; Higher Education; Listening Comprehension Tests; *Listening Skills; *Measurement Techniques; *Speech Skills; Speech Tests; Standards; Test Construction; *Testing Problems; Test Selection
IDENTIFIERS	College Entrance Examination Board

ABSTRACT

The College Board has identified several basic intellectual competencies thought to be essential for effective work in all fields of college study, among them listening and speaking. An issue that arises in connection with these competencies is the availability of suitable measures to assess students' development in these areas. This report considers the availability and adequacy of existing measures of speaking and listening, and discusses a number of issues that should be considered in any efforts to develop new measures of these skills. These issues include: availability and adequacy of existing measures; defining listening and speaking; developing content specifications; relationships among reading, writing, listening, and speaking; instructional effort directed towards speaking and listening; conforming to professionally accepted standards for educational and psychological tests; and administrative feasibility and costs. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249287

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

S. Gardner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

College Board Report

No. 84-5

Considerations for Developing Measures of Speaking and Listening

Donald E. Powers

TM 840 046

Considerations for Developing Measures of Speaking and Listening

Donald E. Powers

Educational Testing Service

College Board Report No. 84-5

ETS RR No. 84-18

College Entrance Examination Board, New York, 1984

Donald E. Powers is Senior Research Scientist with Educational Testing Service, Princeton, New Jersey.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101. The price is $5.

Copyright © 1984 by College Entrance Examination Board. All rights reserved.
The "College Board" and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 1

Issue 1: Availability and Adequacy of Existing Measures 1

Issue 2: Defining *Listening* and *Speaking* 2

Issue 3: Developing Content Specifications 2

Issue 4: Relationships Among Reading, Writing, Listening, and Speaking 3

Issue 5: Instructional Effort Directed Towards Speaking and Listening 4

Issue 6: Conforming to Professionally Accepted Standards for Educational and Psychological Tests 4

Issue 7: Administrative Feasibility and Costs 5

Recommendations 6

References 8

ABSTRACT

The College Board has identified several basic intellectual competencies thought to be essential for effective work in all fields of college study, among them listening and speaking. An issue that arises in connection with these competencies is the availability of suitable measures to assess students' development in these areas.

This report considers the availability and adequacy of existing measures of speaking and listening, and discusses a number of issues that should be considered in any efforts to develop new measures of these skills.

INTRODUCTION

As a result of widespread consultations with hundreds of educators throughout the nation, the College Board's Council on Academic Affairs has identified several broad intellectual skills thought to be essential for effective work in all fields of college study. These basic competencies include reading, writing, speaking and listening, mathematics, reasoning, and studying (The College Board, 1983). A more detailed specification of listening and speaking skills includes the following abilities:

- to engage critically and constructively in the exchange of ideas, particularly during class discussions and conference with instructors;
- to answer and ask questions coherently and concisely, and to follow spoken instructions;
- to identify and comprehend the main and subordinate ideas in lectures and discussions, and to report accurately what others have said;
- to conceive and develop ideas about a topic for the purpose of speaking to a group; to choose and organize related ideas; to present them clearly in standard English; and to evaluate similar presentations by others;
- to vary one's use of spoken language to suit different situations.

Recently, a number of efforts have been undertaken to examine the quality of education in the United States. Among the recommendations made in conjunction with one such effort was the adoption of more rigorous and *measurable* standards by secondary schools, colleges, and universities (National Commission on Excellence in Education, 1983).

Arising naturally both from such recommendations and also from the specification of basic academic competencies are questions about the availability of suitable instrumentation to assess the degree to which students have developed these competencies. A number of standardized tests and other measures, either experimental (or under development) or commercially available, have been designed to assess basic skills in each of the six designated areas, including speaking and listening. However, with respect to the availability of measures, the situation seems less encouraging outside the three traditional basic competencies of reading, mathematics, and writing. Despite some serious gaps in the instruments currently available to assess listening and speaking; however, there appears to be a relatively solid base for further research and development.

The purpose of this report is to take a further step towards the availability of suitable measures of listening and speaking by outlining some of the most critical issues that require attention in any development effort and by suggesting, in a preliminary fashion, some of the characteristic features that any resulting instrument(s) might take.

Issue 1: Availability and Adequacy of Existing Measures

Twenty years ago, Dixon (1964) referred to listening as the most neglected of the language arts and pointed to the lack of adequate tests of listening skills. Kelly's (1965) skepticism about the most widely used listening tests was accompanied by a call for more inventive measurement efforts. Later, Kelly (1966) argued that: "currently published listening tests are not valid measures of a unique skill..." (p. 455). He also argued that: "researchers to date have failed to produce any statistical verifications of listening test validity. In fact, evidence strongly suggests that currently published listening tests measure the same basic factors that are more reliably measured by established achievement tests not involving listening" (p. 458). He concluded that: "A listening test that measures anything but verbal or general mental ability is hard, if not impossible to find" (p. 450).

Larson et al. (1978) repeated essentially this same criticism of listening tests 13 years later, stating that: "None of the current listening tests seem to show adequate levels of validity, and we are not certain that these tests actually measure listening" (p. 57). This conclusion was reached after a careful review of some 90 tests of communication skills, 53 of which were designed for adults or college-level students. The authors also concluded that no one test provided a comprehensive evaluation of speaking *and* listening skills.

In general, then, commercially available tests of listening have often been criticized as being little more than reading tests presented orally instead of tests that tap some unique aspect of the listening process (D.L. Rubin et al., 1982). All tests have not entirely lacked merit, however. Larson et al. (1978), for example, described the STEP Listening Test as "...clearly one of the better-designed and better-normed tests of basic listening skills" (p. 146), but even this test has encountered relatively severe criticism. Lundsteen (1979), for instance, saw this measure as a mix of reading and listening, since, although stimulus passages are presented orally, the multiple-choice options must be read

by examinees. She also hypothesized that many of the test questions could be answered without having heard the oral material. Others (e.g., D.L. Rubin et al., 1982) have also cited the need to read response options as a serious confounding factor in tests of listening. This confounding may explain, at least partially, the substantial correlations of such measures with more general measures of verbal ability (e.g., Anderson and Baldauf, 1963).

A recent local effort by the staff at St. Edward's University has resulted in an audio/videotaped test of listening that has been reviewed quite favorably as "probably the most valid test currently available for use with secondary school students" (Kelley et al., 1983, p. 6). To date, however, relatively little evidence is available regarding the psychometric characteristics of the measure.

With respect to speaking, Bachman and Palmer (1983) characterized the measurement of oral proficiency as "one of the areas of most persistent difficulty in language testing" (p. 154) and pointed to the questionable validity of measures of this construct as well as to the distinctness of the construct. When oral language is assessed, which is seldom, it is generally accomplished through high-inference general impressions or ratings (McCaleb, 1979), which are prone to lack of agreement among judges—a state that D.L. Rubin et al. (1982) describe as "the bane of speech performance rating procedures" (p. 298). Loban (1976) reported finding no published test that attempts to measure the spoken word. After reviewing 71 instruments designed to measure either speaking or listening, K.L. Brown et al. (1979) found no available test that met more than half of the 15 guidelines they determined such measures should meet. It seems clear then that there is no abundance of adequate measures of speaking and listening.

Efforts to institute assessments of speaking and listening have had mixed success at both the national and state levels. Mead (1978) suggested that the impetus for her paper was a "rather disappointing" pilot effort to develop listening measures for the National Assessment of Educational Progress (NAEP), which currently assesses neither speaking nor listening skills.

At the state level, one of the most extensive efforts has been by the Massachusetts Department of Education (1980), which upon initial investigation cited an almost total lack of suitable measures of either listening or speaking. (The most suitable listening measure, it was decided, was the STEP Listening Test.) Recently, other states also have shown considerable interest and activity in assessing listening and speaking skills. Backlund et al. (1982) reported that, as of July 1981, the 50 states differed markedly in their development of assessments of speaking and listening. While 19 states had no communication programs (nor any immediate plans to develop them), 14 reported intentions to develop such programs, although they had not begun to plan specific activities. Four states had identified speaking and/or listening skills and were developing statewide assessment procedures, and 11 states had identified skills for which curriculum materials, *but not assessment procedures*, were under development. Only 4 states had identified skills and had developed (or were developing) both curricula and assessment procedures.

Issue 2: Defining *Listening* and *Speaking*

In summarizing 50 years of research on listening, Devine (1978) concluded that there is no simple, completely acceptable definition of *listening*. This lack of agreement seems to extend to virtually every aspect of the listening process, including its underlying dimensions, methods of assessment, and methods of instruction (Larson et al., 1978). It seems also to apply to communicative competence in general. As D.L. Rubin et al. (1982) point out, there is no consensus about a concept of communicative competence, a situation that has led some (e.g., Backlund et al. 1982) to urge more consistency in the definition and assessment of oral communication skills. This lack of consistency has sometimes resulted in disagreements about the adequacy of existing measures. For example, Kelly (1965) found that the two most widely used listening comprehension tests failed to correlate more highly with each other than with either a reading test or a test of general intelligence and concluded that this evidence constituted good reason for "viewing the best known of the existing listening tests with skepticism." One test author replied that he saw nothing scandalous in the low correlations, but instead saw this merely as confirmation that the tests were built on different theories and definitions (C. Brown, 1967).

The first issue, then, is to come to grips with a suitable operational definition, if not a theory, of the particular communication skills to be assessed. As D.L. Rubin et al. (1982) have suggested: "the lack of conceptual clarity may be the greatest impediment facing communication measurement" (p. 296).

We assume that any operational definition will, at a minimum, be guided by the Board's specification of basic listening and speaking competencies, and also by other considerations discussed below.

Issue 3: Developing Content Specifications

Despite the lack of agreement on definitions of listening and speaking, or perhaps as a consequence of it, taxonomies of the roles, functions, and settings of communication abound (Rubin, 1981). Examples can be found in Allen and Brown (1976), Bassett et al. (1978), Joos (1961), Levison (1976), Richards (1983), and Wiemann and Backlund (1980). While the College Board's statement of basic competencies is of paramount importance, any test development effort in this area will also need to be informed by various other taxonomies and will require procedures for accommodating them.

Delineating the content domain may be easier for most tests of educational achievement than for measures of

speaking and listening skills, for which the typical content specification procedures may be insufficient. D.L. Rubin et al. (1982) suggest that traditional methods may not work because (1) oral communication instruction is not well-established in educational curricula and, therefore, tests of listening and speaking may measure skills that schools do not undertake to teach and, (2) as stated previously, no consensus exists about a concept of communicative competence that could reasonably guide test construction. Furthermore, because of the variety of objectives, skills, and lists of competencies that are available, methods for determining priorities and sampling taxonomies will be required. Undoubtedly, it will not be feasible to assess comprehensively the entire domain of listening and speaking abilites.

Mead (1978) has provided a useful start for identifying the listening objectives "most often identified in instructional and assessment material":

1. to recall significant details;
2. to comprehend main ideas;
3. to draw inferences about information;
4. to make judgments concerning the speaker (e.g., attitude, intent);
5. to make judgments about the information (e.g., type of evidence, logic of arguments).

One attractive possibility for narrowing the measurement of these competencies is to concentrate on the limited category of speaking and listening skills referred to as "informing" communication acts (referential communication), which focus on "communication effectiveness" or "the ability to inform with accuracy and efficiency" (Dickson and Patterson, 1981). More will be said later about this approach.

Issue 4: Relationships Among Reading, Writing, Listening, and Speaking

Whether language proficiency can be partitioned into distinct components that can be taught and tested separately or whether it is essentially a single global trait is a question that has received a good deal of attention (see, for example, Palmer et al. 1981; Oller and Hinofotis, 1980; Scholz et al. 1980). The evidence appears to be somewhat inconclusive, however, with some studies supporting each of the two options. Most research does reveal substantial correlations among measures of listening, speaking, reading, and writing, however, which is not surprising, since many of the same processes underlie each skill. For example, as Daneman et al. (1982) pointed out, except for visual decoding, both listening and reading appear to involve similar processes, which include speed of lexical access, efficiency in semantically integrating information, and retention of information in working memory.

Evidence of the substantial correlations between listening and speaking and other more general verbal abilities comes from several sources, including studies of English as a second language. For example, Swinton and Powers (1980) reported average intercorrelations for six forms of the Test of English as a Foreign Language (TOEFL) as follows: .70 between listening comprehension and structure/written expression and .69 between listening comprehension and vocabulary/reading comprehension. Yet, even with these high correlations, Swinton and Powers (1980) found clear evidence for a distinct listening comprehension factor in their analyses.

In exploring the development of a speaking test for foreign students, Clark and Swinton (1979) reported correlations of .58, .61, and .65 between a measure of oral proficiency (The Foreign Service Interview) and each of the three sections of the TOEFL (which does not contain an oral measure). Bachman and Palmer (1981) also reported high correlations for foreign students between various measures of speaking and reading. The correlations of five different methods of measuring reading with the corresponding methods of measuring speaking ranged from .46 to .69, with a median of .64.

Similarly, high correlations among verbal measures also have been reported for English-speaking students. For example, Devine (1967) summarized a number of studies that show correlations of .58 to .82 between measures of listening and measures of general or verbal intelligence. A relatively large number of studies can be cited, most of which document the fairly substantial correlations between various measures of listening or speaking and other verbal measures.

Although these skills are highly interrelated, they usually are assumed to be distinct, at least logically (Hutson, 1982). However, because the intercorrelations among measures of these skills are typically so high, the question remains regarding the practical significance of separately measuring each skill.

Lundsteen (1979) has suggested that, "Listening and speaking have been considered the base for other language skills" (p. xi). In fact, she points out that reading may depend so completely on listening ability that it appears to be a special extension of reading. This is perhaps not surprising, since listening is the very first language skill to develop, followed in order by speaking, reading, and writing.

There are some obvious similarities among the four aspects of language proficiency. Reading and listening can be described as receptive skills, while speaking and writing are productive in nature. The common code of listening and speaking is one of sound, while that of reading and writing is one of print. In many ways, the relationships between listening and speaking parallel those between reading and writing, in particular the relationship of the listener to the speaker and the reader to the writer. Many of the issues in listening (or speaking, reading, or writing) can be examined in terms of the interaction of the listener (or speaker, etc.)

and the tasks they perform (Hutson, 1982).

While there are many similarities among the four language skills, there are also some important differences, for example, in their situational contexts. The reader, for example, is usually alone and unable to interact with the writer, whereas the listener can typically interact by questioning the speaker. The listener can also study facial expressions and other nonverbal language and can pick up signals from the stress, pitch, etc., of the speaker's voice. Spoken language is different from written language in that it is nonlinear, incomplete, and redundant (Mead, 1978). The listener therefore, being "caught in the flow of time," must rely to a greater degree on memory, whereas the reader can reread material, thereby checking his or her understanding.

The point here is that there are both similarities and differences among the four skills that must be considered in the development of listening and speaking measures. It will not suffice simply to administer reading passages orally in order to assess listening comprehension. The issue, therefore, is to apply what is known about listening, speaking, reading, and writing towards implementing Mead's (1978) recommendation to focus on the skills that are unique and central to listening (and speaking). This applies especially to their interactive nature.

Issue 5: Instructional Effort Directed Towards Speaking and Listening

A potential use of any measures of speaking and listening would presumably be to determine if what is taught is being learned. Yet, as mentioned above, oral communication instruction is generally not well-established in the educational curricula of American secondary schools (D.L. Rubin et al., 1982), and when communication instruction is undertaken, it most often focuses on reading and writing (Bassett et al., 1978). Although listening instruction appears to be receiving somewhat greater attention than in prior years, "it still remains the 'orphan' of the language arts" (Wolvin and Coakley, 1979, p. 1) and, therefore, few students have received formal instruction in listening. Likewise, few American high school graduates have received any instruction in oral communication. Recent data from the National Commission on Educational Statistics (NCES) show that fewer than 65 percent of secondary schools offer speech communication courses (Lieb-Brilhart, 1980), most of which are probably electives taken by relatively few students. As noted above, relatively few states have formal communication programs (Backlund et al., 1982), although a handful of states, most notably Massachusetts (Brown et al., 1979), have well-developed programs.

Thus, the issue to be addressed here concerns the nature of a testing effort that may relate only indirectly to what is taught (although it may relate more directly to what is learned). In this regard, there may be considerable potential for influencing what is taught with respect to speaking and listening. All of this points to a need for measures of speaking and listening that would, at the least, encourage instruction in these skills and, hopefully, facilitate such instruction.

One noteworthy example of an effort to develop both curricular and assessment materials is the work of the staff at St. Edward's University in Austin, Texas. A listening skills program, designed for underprepared Hispanic secondary school students, has been incorporated successfully into classroom instruction. A visiting team has reported that teachers have reacted favorably to the materials and have reported improvements in students' listening skills (Kelley et al., 1983).

Issue 6: Conforming to Professionally Accepted Standards for Educational and Psychological Tests

It is expected that any tests of listening and speaking would conform to the commonly accepted technical standards that apply to all professionally developed tests. Besides meeting the existing standards (American Psychological Association, 1974), efforts to devise new measures of speaking and listening should anticipate and meet the yet to be released revisions of these standards (AERA, APA, NCME, 1983), which present requirements for sound professional practice. Some of these requirements—for example, to demonstrate construct validity—occupy a central place in the acceptance of these tests, both by the professional measurement community and by potential test users. Because of nagging questions about what such previous tests have actually measured, it seems of paramount importance to provide adequate documentation of construct validity through carefully designed research studies.

Other issues, for example, reliability, are equally important to address because of the well-known problems with speech rating scales. Considerable progress has been made in assessing writing, with respect to identifying sources of rater error and in devising methods to enhance inter-rater reliability, but the "lack of agreement among judges is the bane of speech performance rating procedures" (D.L. Rubin et al., 1982). A very useful and comprehensive assessment of rating procedures is available to guide future efforts (Saal et al., 1980), as is a discussion of speech rating methods (Bock and Bock, 1981).

Undoubtedly, any test development efforts will need to address such issues as the number of speech samples and the number of raters required to produce adequately reliable assessments. There is some reason for optimism on this front. For instance, Marine (1965) estimated the reliability of ratings for a single speech segment to be .78 and for three speech segments .91. D.L. Rubin (1981) has implied that these results may suggest the need for only a single sample of speech. Clearly, however, a thorough assessment of factors contributing to test reliability would be needed.

Reliably measuring listening may be more problematic

than measuring speaking, however, because listening performance is so dependent on attentional processes, and therefore is moderated by temporary (and extraneous) characteristics of test takers, which may substantially affect reliability (D.L. Rubin et al., 1982). Earlier, Kelly (1965) also pointed to the "great vulnerability" of listeners to distractions such as fatigue, lack of motivation, personal worries, antagonism towards the speaker, disinterest in the topic, momentary daydreams, and outside interferences—a considerable list! Moreover, in comparison with the reader, the listener may be especially prone to such distractions because he or she cannot retrace his or her steps. This state of affairs led Kelly (1965) to surmise that "perhaps listening tests will always be less reliable than most other types of measuring devices..." (p. 143).

In addition, tests of language proficiency may be more susceptible than other kinds of measures to possible unfairness resulting from particular kinds of test content. Mead (1978) found, for instance, the "surprising result" that for listening items, unlike other communication items being tried out for the National Assessment of Educational Progress (NAEP), a high proportion were correlated with ethnicity. She recommended that extraneous factors such as the dialect used by both speakers and listeners be considered as a possible source of test bias. With respect to speech, some relevant points have been made by Freedle (1981), who discussed some breakdowns in communication that result from cultural differences. For example, it has been reported that Navajo Indians prefer not to comment on a topic unless they regard themselves as highly proficient in it; to speak prematurely is considered a breach of intelligent behavior. D.L. Rubin et al. (1982) also have suggested that communication testing itself may present an anomalous situation for some individuals because of previous socialization. Norton and Hodgson (1974) also presented some interesting information on the intelligibility of black and white speakers to black and white listeners. Stiggins (1981) has discussed a number of potential sources of bias in listening and speaking tests.

Thus while all of the usual psychometric standards pertain to tests of speaking and listening, at least three issues—validity, reliability, and test bias—may be especially critical for measures of speaking and listening, and therefore may deserve even more careful consideration.

In addition to the standards that apply generally to all standardized tests, there are also those that pertain more specifically to measures of listening and speaking. Among these are the guidelines promulgated by state education agencies, for example, the comprehensive set provided by K.L. Brown et al. (1979) in connection with the Massachusetts statewide assessment. These guidelines have subsequently been endorsed by the Speech Communication Association (SCA). The 15 criteria presented by the SCA pertain to both content specifications (e.g., instruments should emphasize application of speaking and listening skills in familiar situations) and technical specifications (e.g., assessments should be consistent with other available evidence). It may not be possible, or perhaps even entirely appropriate, to meet each of these SCA criteria in any future development efforts, but they should be carefully considered.

Issue 7: Administrative Feasibility and Costs

One of the major problems confronted in the measurement of communication skills is administrative feasibility. Because, as D.L. Rubin, et al. (1982) contended, communication is essentially a social act, tests of communicative competence are likely to be more complex, and therefore more time-consuming and expensive, than other kinds of large-scale assessments. Also, the give and take of typical communication acts is difficult to recreate in a test setting (Mead, 1978). Most measures of oral communication skills, for example, typically require many hours of assessment for a relatively small number of examinees, a situation that is in direct contrast to secondary school needs to assess large numbers of students in a limited amount of time (McCaleb, 1979). The assessment of listening skills is complicated because listening is a covert act and, hence, there is no directly observable product to evaluate. Speaking competence, unlike certain other modes of communication like writing, has proved remarkably resistant to indirect measurement (D.L. Rubin, 1981), that is, to assessing knowledge about the conventions of speaking. Therefore, speech performance ratings, which require the use of trained raters, as well as both individual administration and individual evaluation procedures, appear to be the most likely means of assessing speech. If the substantial time needed to obtain, listen to, and rate students' speech is diverted from instruction, the testing of communication skills may meet with significant resistance from teachers, who often regard large-scale assessments with suspicion, especially if such assessments do not directly support instruction.

Nonetheless, even with these problems, several examples attest to the fact that large-scale assessments of speaking and listening are possible. In 1957, the College Board began to offer listening comprehension tests to assess achievement in several foreign languages. The Board also has continued to offer measures in both speaking and listening skills in selected foreign languages as part of its Advanced Placement (AP) Program. In recent months, the AP Program has successfully experimented with assessing German speaking proficiency over the telephone (Haag, 1983). In this approach, examinees give their responses directly to Educational Testing Service (ETS) raters via phone lines from their schools, where they can be observed by test administrators. The program appears workable for the German test because of the relatively small number of test takers involved (when compared with the French and Spanish examinations, for which examinee responses are tape recorded). However, the use of telephone assessments with large numbers of candidates is thought to pose potential test security problems.

Since its inception, the TOEFL program has also included a section on listening comprehension in its test, and several years ago introduced the Test of Spoken English (TSE) after a rather extensive research and development effort (Clark and Swinton, 1979; 1980). The TSE development effort is noteworthy not only because an operational measure eventually emerged, but also because the project reports contain examples of item types that, although not included in the final operational version of the TSE, could provide a useful departure for future test development. As reported in the *Manual for Score Users* (ETS, 1982), the TSE contains several kinds of tasks, each involving a particular speech activity ranging from reading a printed passage aloud to answering a series of spoken questions designed to elicit free and somewhat lengthy responses. The test is administered with cassette or reel-to-reel tape recorders using a multiple-recording facility such as a language laboratory. The answer tapes then are judged independently by two raters who are previously trained at one-day workshops. If ratings do not agree, a third rater resolves discrepancies.

According to TOEFL program staff, the first several years of the TSE program have involved relatively small numbers of examinees each year, and the program's initial years have required a rather substantial subsidy. It is hoped, however, that a break-even point might be reached in the near future.

The College Outcomes Measurement Project (COMP) of the American College Testing Program (ACT) provides both listening and speaking measures as part of an extensive battery (ACT, 1979). Although the entire battery is time-consuming to administer and score, provisions are made for selecting particular areas and components to fit the needs of individual institutions. Also, long, open-ended versions, as well as shorter, multiple-choice versions of the speaking measures are available. It is not totally clear, however, that these versions are actually measuring the same competencies.

Finally, several states also have demonstrated the feasibility of large-scale assessments of speaking and listening. The Massachusetts Department of Education (1980) assesses twelfth-grade students' listening skills with six tape recordings of common listening situations and associated multiple-choice questions. Speaking is assessed by two methods: (1) each student's speaking is observed and rated by two teachers in whose classes the student is enrolled, and (2) students are evaluated on a one-to-one basis by specially trained evaluators.

With respect to standards and guidelines, R.B. Rubin (1982) has demonstrated the feasibility of meeting guidelines by developing the Communication Competency Assessment Instrument, which she believes meets all of the criteria espoused by the Speech Communication Association. This particular instrument may bear careful inspection.

RECOMMENDATIONS

Listening and speaking are intimately related; even with this intimacy, however, procedures that do not artificially isolate speaking from listening are "conspicuous in their rarity" (D.L. Rubin et al., 1982, p. 30). The most interesting examples of exceptions to this come from the research on referential, or "informing," communication, which offers an attractive possibility for integrating the measurement of speaking and listening skills—a course of action that we are recommending not only for purposes of valid measurement but also for administrative efficiency. This course is also consistent with the current emphasis in language research and testing on functional communicative competence.

Dickson and Patterson (1981) described the gist of referential communication as attempting to communicate to another person about a target referent. Accuracy is defined as how well the listener understands the communication, as evidenced by his or her ability to identify the referent among a set of alternatives or to reproduce it with some degree of fidelity. The procedure may involve interaction and questioning by the listener. Dickson and Patterson (1981) listed several types of tasks that have been used—picture choosing, placing in an array, map directions, and model building. In picture choosing, the speaker must tell the listener how to choose a specific picture from several alternatives. Placing in an array is a variant of picture choosing in which the speaker tells the listener how to place objects in an array, thus increasing the task demands by introducing the need to communicate about location and orientation. Dickson and Patterson (1981) found that the referent sets used in conjunction with research using these types of tasks have been "rather unimaginative," being restricted largely to attributes such as color, shape, and size. They recommended the development of more diverse, and interesting, attributes, a suggestion with which we concur.

The map directions task involves giving directions to a listener after studying a map. The demands, and thus difficulty, of the task can be increased by, for example, increasing the number of details on the map and the complexity of the route to be followed. Model building involves activity such as explaining to the listener how to build a model that is identical to an already assembled model given to the speaker. This task is thought to be especially worthy of consideration if the objective is to assess the interactive nature of communication. Baldwin and Garvey (1973) provided an account of some research with fifth-grade students doing these types of tasks and suggest a procedure to manipulate systematically the attributes of the target referents.

As stated previously, most referential communication tasks, besides using rather unimaginative stimuli, reflect a relatively limited category of listening and speaking skills. Many of these tasks are suitable primarily for younger

students, although some tasks have been developed for older students. For example, Lieb-Brilhart (1965) developed a task in which college undergraduates were asked to describe a preassigned geometric design to an audience, whose members were asked to duplicate the figure as described. Each individual served as a speaker once and, in turn, as a listener for every other member of the audience. Each speaker was scored according to how well listeners duplicated the figure described, and each listener was scored according to how well the figures were duplicated by each speaker. Lieb-Brilhart (1965) found that scores on this measure were related to scores on one of the most widely used listening measures, the Brown-Carlsen Listening Test, but that the separate listening and speaking scores derived from the measure were uncorrelated. It is recommended that the extension of this model be explored, employing, for example, tasks that reflect more closely the general communication demands of academic settings and, specifically, the Board-designated basic competencies in speaking and listening. R.B. Rubin's (1982) extension to an education context of Bassett et al.'s (1978) three-context (occupational, citizenship, and maintenance) communication taxonomy should also prove to be helpful in this regard, since this extension specifies competencies needed for communication with instructors, fellow students, academic advisors, and so on. Some promising evidence of concurrent validity was reported by R.B. Rubin (1982), who found a correlation of .69 between Lieb-Brilhart's (1965) earlier measure and her Communicative Competency Assessment Instrument, which is based on the extension.

Besides extending the referential communication model to other tasks involving the description of target referents, we suggest that a more general model in which students act as both the evaluator and the evaluatee be considered. In this approach, speakers would be rated by their peers as to how well they accomplished any number of communicative tasks involving, for example, informing, persuading, and so on. Students would first be trained with respect to the criteria on which they would base their evaluations. Students, as listeners, could be asked to provide judgments of how well various speaking functions were accomplished by their fellow students. Although these judgments probably cannot be scored as reliably as the other kinds of referential communication tasks mentioned above, we expect that the availability of an entire classroom of student raters would produce quite reliable ratings. As with other tasks, listeners could also receive scores according to how well their ratings correspond with the "truth," possibly as described by a classroom teacher or by the class itself (for example, the average rating made by the entire class). More efficient alternative models might also be explored, for example, having students work in pairs or in small groups, periodically changing partners. Basically, this general strategy conforms to Bock and Bock's (1981) suggestion that students should be trained to become good evaluators (listeners): "If communication classes are to train good listeners as well as good speakers, *peer evaluation* becomes essential" (p.18).

Lammlein and Borman (1979) present a thorough review of research on peer-rating procedures, which could be consulted for further development work. Of course, there are numerous kinds of rating scales from which to choose, several of which (e.g., general impression or holistic marking, analytic scales, dichotomous choice scales, and rhetorical trait scales) apply directly to assessing speaking and listening (D.L. Rubin, 1981).

In summary, what is suggested is a system of assessment, more than a traditional standardized test. Development would entail (1) a variety of communication stimuli, (2) suggested procedures for administering the system, and (3) a method for systematically evaluating and recording performances. It is quite possible that advances in computer-assisted testing and other technologies could facilitate the administration of this system. Videodisk and audiodisk technology might play a significant role. At present, however, we are unaware of any computer-based system for assessing speaking and listening skills.

Finally, we offer the following recommendations as guides for any future development of measures of listening and speaking:

1. Specify from existing taxonomies, from surveys of instructional activities, and from other sources the listening and speaking skills that correspond most closely with the Board-specified competencies under speaking and listening.

2. Model the development of a listening/speaking assessment instrument (or system) after instruments used in so-called referential communication tasks, but extend previous work to include more educationally relevant tasks.

3. Take advantage of the reciprocal relationship between listening and speaking, focusing on the unique interactive nature of speaking/listening communication.

4. Employ students in the evaluation and scoring of speaking performance, utilizing students' abilities to recognize good speaking, perhaps after suitable training.

5. Pay special attention to the psychometric requirements for reliability, validity (especially construct validity), and test fairness, possibly by specifying research activities to investigate these requirements.

REFERENCES

Allen, R. R., Brown, K. L. (Eds.). *Developmental communication competence in children*. Skokie, Ill.: National Textbook, 1976.

American College Testing Program. *The college outcome measures project of the American College Testing Program (COMP prospectus)*. Iowa City: Author, 1979.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Joint technical standards for educational and psychological testing* (draft report). Authors, February 1983.

American Psychological Association. *Standards for educational and psychological tests*. Washington, D.C.: Author, 1974.

Anderson, H. M., and Baldauf, R. J. A study of a measure of listening. *Journal of Educational Research*, 1963, *57*, 197–200.

Bachman, L. F., and Palmer, A. S. The construct validity of the FSI oral interview. *Language Learning*, 1981, *31*, 67–86.

Bachman, L. F., and Palmer, A. S. The construct validity of the FSI oral interview. In J.W. Oller (Ed.), *Issues in language testing research*. Rowley, Mass.: Newbury House, 1983.

Backlund, P., Booth, J., Moore, M., et al. *State practices in speaking and listening skill assessment*. Annandale, Va.: Speech Communication Association, 1982.

Baldwin, T. L., and Garvey, C. J. Components of accurate problem-solving communications. *American Educational Research Journal*, 1973, *10*, 39–48.

Bassett, R. E., Whittington, N., and Staton-Spicer, A. The basics in speaking and listening for high school graduates: What should be assessed? *Communication Education*, 1978, *27*, 293–303.

Bock, D. G., and Bock E. H. *Theory & research into practice: Evaluating classroom speaking*. Urbana, Ill.: ERIC Clearinghouse on Reading and Communication Skills; and Annandale, Va.: Speech Communication Association, 1981.

Brown, C. Response to Kelly: Special report. *Speech Monographs*, 1967, *34*, 465–466.

Brown, K. L., Backlund, P., Gurry, J., and Jandt, F. *Assignment of basic speaking and listening skills* (2 vols.). Boston: Massachusetts Department of Education Bureau of Research and Assessment, 1979.

Clark, J. L. D., and Swinton, S. S. *An exploration of speaking proficiency measures in the TOEFL context* (TOEFL Research Report No., 4). Princeton, N.J.: Educational Testing Service, 1979.

Clark, J. L. D., and Swinton, S. S. *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report No. 7). Princeton, N.J.: Educational Testing Service, 1980.

College Board. *Academic preparation for college: What students need to know and be able to do*. New York: Author, 1983.

Daneman, M., Carpenter, P. A., and Just, M. A. Cognitive processes and reading skills. In B. A. Hutson (Ed.), *Advances in reading/language research: A research annual* (vol. 1). Greenwich, Conn.: Jai Press, 1982.

Devine, T. G. Listening. *Review of Educational Research*, 1967, *37*, 152–158.

Devine, T. G. Listening: What do we know after fifty years of research and theorizing? *Journal of Reading*, 1978, *21*, 296–304.

Dickson, W. P., and Patterson, J. H. Evaluating referential communication games for teaching speaking and listening skills. *Communication Education*, 1981, *30*, 11–21.

Dixon, N. R. Listening: Most neglected of the language arts. *Elementary English*, 1964, *41*, 285–288.

Educational Testing Service. *Test of Spoken English: Manual for score users*. Princeton, N.J.: Educational Testing Service, 1982.

Froedle, R. O. Interaction of language use with ethnography and cognition. In J. Harvey (Ed.), *Cognition, social behavior, and the environment*. Hillsdale, N. J.: Erlbaum, 1981.

Haag, C. Personal Communication, 1983.

Hutson, B. A. The scope of reading/language research. In B. A. Hutson (Ed.), *Advances in reading/language research: A research annual* (vol. 1). Greenwich, Conn.: Jai Press, 1982.

Joos, M. *The five clocks*. New York: Harcourt, Brace & World, 1961.

Kelley, H. P., Lyle, M. R., Nichols, R. G., and Stokes, V. *Report and recommendations of the St. Edward's University listening skills programs visitation team*. Report to the College Board, July 1983.

Kelly, C. M. An investigation of the construct validity of two commercially published listening tests. *Speech Monographs*, 1965, *32*, 139–143.

Kelly, C. M. Listening: Complex of activities—and a unitary skill? *Speech Monographs*, 1956, *34*, 455–465.

Lammlein, S. E., and Borman, W. C. *Peer rating research: Annotated bibliography (AFHRL-TR-79-9)*. Brooks Air Force Base, Texas: Personnel Research Division, Air Force Human Resources Laboratory, June 1979.

Larson, C., Backlund, P., Redmond, M., and Barbour, A. *Assessing functional communication*. Urbana, Ill.: ERIC Clearinghouse on Reading and Communication Skills; and Falls Church, Va.: Speech Communication Association, 1978.

Levison, G. K. The basic speech communication course: Establishing minimal oral competencies and exemption procedures. *Communication Education*, 1976, *25*, 222–230.

Lieb-Brilhart, B. The relationship between some aspects of communicative speaking and communicative listening. *Journal of Communication*, 1965, *15*, 35–46.

Lieb-Brilhart, B. Effective oral communication programs: Myths and tensions. In *Resources for Assessment in Communication*. Annandale, Va.: Speech Communication Association, 1980.

Loban, W. Language development and its evaluation. In A. H. Grommon (Ed.), *Reviews of selected published tests in English*. Urbana, Ill.: National Council of Teachers of English, 1976.

Lundsteen, S. W. *Listening: Its impact at all levels on reading and the other language arts* (rev. ed.). Urbana, Ill.: National Council of Teachers of English, 1979.

Marine, D. R. An investigation of intra-speaker reliability. *Speech Teacher*, 1965, *14*, 128–131.

Massachusetts Department of Education. *Massachusetts assessment of basic skills 1979–80, summary report: Speaking and listening*. Boston: Author, September 1980.

McCaleb, J. L. Measuring oral communication. *English Education*, 1979, *11*, 43–47.

Mead, N. A. *Issues related to assessing listening ability*. Paper presented at the American Educational Research Association

meeting, Toronto, Canada, March 1978.

National Commission on Excellence in Education. *A nation at risk: The imperative for educational reform*. Washington, D.C.: U.S. Department of Education, 1983.

Norton, D., and Hodgson, W. R. Intelligibility of black and white speakers for black and white listeners. *Language and Speech*, 1974, *16*, 207–210.

Oller, J. W., and Hinofotis, F. B. Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. W. Oller and K. Perkins (Eds.), *Research in language testing*. Rowley, Mass.: Newbury House, 1980.

Palmer, A. S., Groot, P. J. M., and Trosper, G. A. (Eds.). *The construct validation of tests of communicative competence*. Washington, D.C.: Teachers of English to Speakers of Other Languages 1981.

Richards, J. C. Listening comprehension: Approach, design, procedure. TESOL *Quarterly*, 1983, *17*, 219–240.

Rubin, D. L. *Using performance rating scales in large scale assessments of oral communication proficiency*. Paper presented at the American Educational Research Association meeting, Los Angeles, April 1981.

Rubin, D. L., Daly, J., McCroskey, J. C., and Mead, N. A. A review and critique of procedures for assessing speaking and listening skills among preschool through grade twelve students. *Communication Education*, 1982, *31*, 285–303.

Rubin, R. B. *Assessment of college level speaking and listening skills*. Paper presented at the American Educational Research Association meeting, Los Angeles, April 1981.

Rubin, R. B. Assessing speaking and listening competence at the college level: The communication competency assessment instrument. *Communication Education*, 1982, *31*, 19–32.

Saal, F. E., Downey, R. G., and Lahey, M. A. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 1980, *88*, 413–428.

Scholz, G , Hendricks, D., Spurling, R., et al. Is language ability divisible or unitary: A factor analysis of 22 English language proficiency tests. In J. W. Oller and K. Perkins (Eds.), *Research in language testing*. Rowley, Mass.: Newbury House, 1980.

Stiggins, R. J. Potential sources of bias in speaking and listening assessment. In R. J. Stiggins (Ed.), *Perspective on the assessment of speaking and listening skills for the 1980's*. Portland: Northwest Regional Educational Laboratory, 1981.

Swinton, S. S., and Powers, D. E. *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Report No. 6). Princeton, N.J.: Educational Testing Service, 1980.

Wiemann, J. M., and Backlund, P. Current theory and research in communicative competence. *Review of Educational Research*, 1980, *50*, 185–199.

Welvin, A. D., and Coakley, C. G. *Theory & research into practice: Listening instruction*. Urbana, Ill.: ERIC Clearinghouse on Reading and Communication Skills; and Falls Church, Va.: Speech Communication Association, 1979.